EVALUATION OF THE RISK OF DEATH FOR VERY LOW BIRTHWEIGHT BABIES AND COMPARARISON BETWEEN NEONATAL INTENSIVE CARE UNITS: APLLICATION OF ROC CURVES

ANA C. BRAGA[1], P. OLIVEIRA[1], A. GOMES[2]

[1]University of Minho, Braga, Portugal

[2]Hospital Garcia de Orta, Almada Portugal

ABSTRACT

In general, the neonatal rates of mortality are considered the most important measure for the evaluation of health services and the development of society. Babies with very low birthweight (<1500g at birth) contribute significantly to the increase of the mortality and morbidity rates.

This study aims to compare, in the same set of babies, 4 different indexes of clinical seriousness; these indexes quantify the risk of death for babies with very low birthweight.

The study was conducted on 169 new-borns with birthweight under 1500g, admitted to the Neonatology Unit of the Hospital Garcia de Orta in Portugal, between January 1992 and July 1995. The data were collected retrospectively on the same set of new-borns to permit the comparison of the different indexes.

To compare Neonatal Intensive Care Units, we use another set of new-borns. These data (234 new-borns with very low birthweight) were collected in 4 different Portuguese hospitals (H_1, H_2, H_3 and H_4) during 1995.

The 4 indexes studied were: CRIB (Clinical Risk Index for Babies), TISS (Neonatal Therapeutical Intervention Score System), SNAP (Score for Neonatal Acute Physiology) and SNAP-PE (Score for

Neonatal Acute Physiology-Perinatal Extension). Also, included in the study was the variable WEIGHT (birthweight) because it is an important index to evaluate the initial risk of death.
The selection of the best diagnostic test was based on the Receiver Operating Characteristic (ROC) curves.
The results, based on the ROC curves, allow the conclusion that CRIB is the best index for evaluating the risk of neonatal death, and that hospital H_4 has better performance in terms of neonatal intensive care.

Key Words: Receiver Operator Characteristic curve, ROC Analysis, low birthweight babies.

1. INTRODUCTION

New-born weight has been used as a measure of the neonatal risk, mostly due to its importance and evaluation easiness. However, the need for more accurate forms of evaluation has arisen, thus allowing the comparison between services, regions and countries.
In recent years clinical indexes have been developed with this aim. From those indexes, the most widely used are: CRIB (Clinical Risk Index for Babies), TISS (Neonatal Therapeutical Intervention Score System), SNAP (Score for Neonatal Acute Physiology) and SNAP-PE (Score for Neonatal Acute Physiology - Perinatal Extension). It should be noted that these different scoring systems require the collection of data for different time periods. Thus, for the above mentioned systems, the number of variables to collect is 6 (CRIB), 26 (SNAP), 29 (SNAP-PE) and 48 (TISS). All these variables are collected in the first 24 hours of life, and exceptionally for CRIB, this period is reduced to 12 hours after birth. For this reason, CRIB is the easiest index to be used, either in terms of variables, either in terms of time of collection.
The International Neonatal Network [1], refers that is possible to compare Intensive Care Units, using the index CRIB as a performance measure.
The diagnostics is an imperfect process. A diagnostic test aims to classify individuals into two classes: normal and abnormal; for that reason, there exists always the possibility of making one of two types of errors: to classify an abnormal individual as normal and, vice-versa, to classify a normal individual as abnormal.
The ROC analysis gives a precision index for the diagnostic which is independent from other information, decision factors and priori probabilities. It emerged from the signal detection theory, and was extensively applied in perceptive cognitive studies in psychology, and currently is being applied in other fields, namely medicine.

2. PROBLEM UNDER STUDY

The range of observed values for the various scales results in a overlap of the two distributions of the so called normal and abnormal cases. Graphically, this can be illustrated by Figure 1.

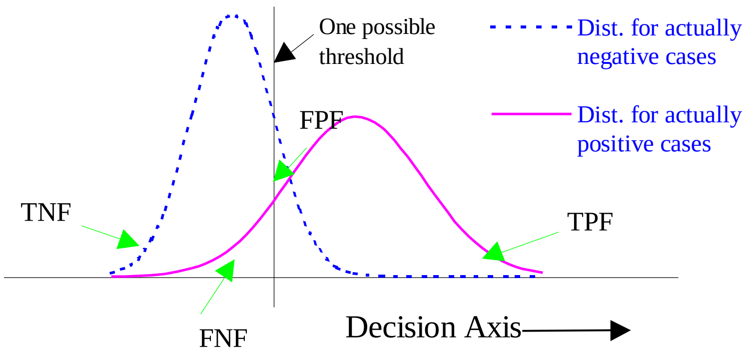


Fig. 1 - Two hypotetical distributions of a quantity on wich a decision is based, showing one possible threshold, Metz [2].

For any diagnostic test a cut-off value is fixed for the variable under study; this value determines the classification of the individuals as abnormal and normal. Thus, any test is evaluated by the relative comparison of the fractions of true positive (TPF), false positive (FPF), true negative (TNF) and false negative (FNF).

Figure 1 explicits the relation between the cut-off value and the definition of these fractions; clearly, reducing the FPF implies an increase on the FNF. In general, a test is evaluated by two measures, sensitivity (TPF) and specificity (TNF). For different cut-off values it is possible to obtain different pairs of TPF and FPF. These pairs of values can be represented as values in a co-ordinate system, known as ROC curve.

3. DIAGNOSTIC TESTS

In the diagnostic test we have as hypothesis for the problem:

H_0: The new-born is going to die, D

H_1: The new-born is going to survive, S

In practice, it is desirable to have a test that is at the same time highly sensitive and highly specific. Bearing in mind that the cut-off value defines the rejection region, the ROC analysis produces the probability of accepting H_0, i.e., to assume that the new-born of very low weight has a high risk of dying.

4. METHODOLOGY

A general method to test if the difference between two areas beneath the ROC curves from the same set of patients is significant is based on the z critical ratio defined by Hanley and MacNeil [3].
This z quantity reports to the standard normal tables, and z values above a given value, imply that the two true ROC areas are significantly different.
The areas under the ROC curves and the standard errors associated with these areas can be obtained in different ways, as it is reported on Hanley and MacNeil [3]; in this work, the Wilcoxon-Mann-Whitney statistic has been used to compute the areas, being the standard errors associated to the areas obtained from the variance of the Wilcoxon statistic.

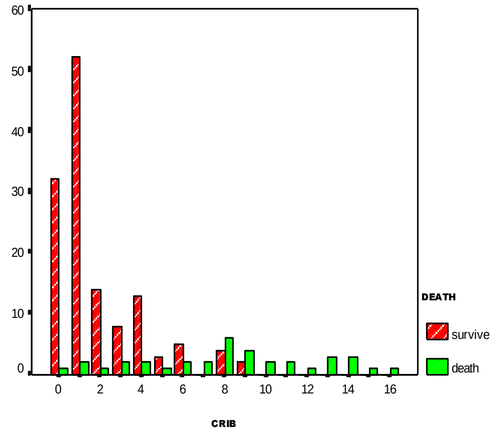
When the data are from the same set of individuals, the samples cannot be considered independent, and thus, it is necessary to compute the correlation coefficient between the areas, as it is reported on Hanley and MacNeil [3].

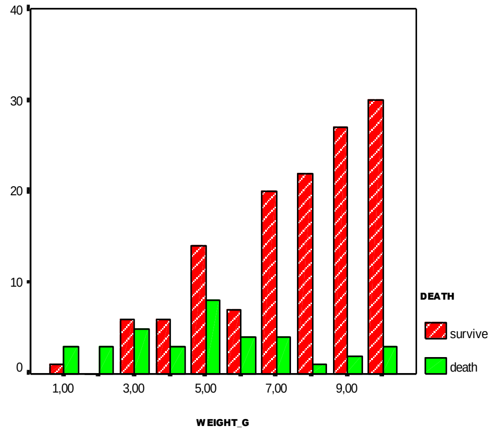
5. DATA DESCRIPTION

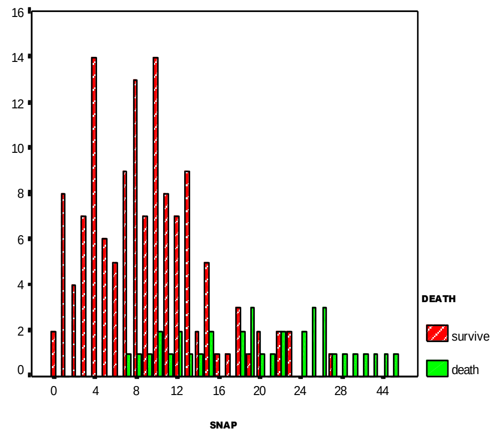
The set of individuals under study is formed by 169 new-borns of very low weight (under 1500 grams) admitted to the Neonatology Intensive Care Unit of the Hospital Garcia de Orta in Portugal. The data were collected retrospectively over the same set, in order to allow the comparison between the different scales, during 3 years, ranging from 1992 to 1995. From these 169 new-borns, 133 survived, having been observed 36 deaths. The indexes evaluated were (CRIB), (SNAP), (SNAP-PE) and (TISS). An extra index was formed, WEIGHT_G, that corresponds to the grouping of the continuous variable weight, into 9 classes, ranging between 540 g and 1500g.
To compare Neonatal Intensive Care Units, we use another set of new-borns. These data (234 new-borns with very low birthweight) were collected in 4 different Portuguese hospitals (H_1-77, H_2-33 , H_3-45 and H_4-79) during 1995.

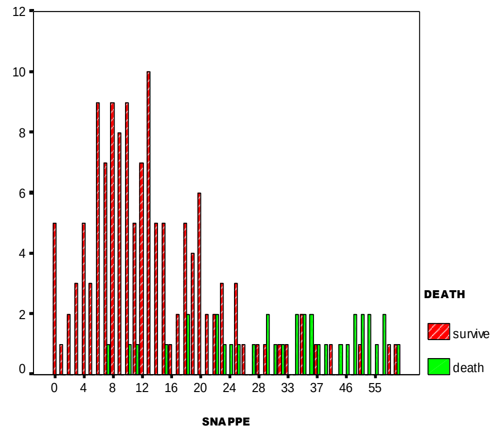
6. EXPERIMENTAL RESULTS

Figure 2(a) presents the frequency distributions for the new-borns that survived and for those that passed away for the different indexes, and figure 2(b) presents the frequency distributions for the new-borns that survived and for those that passed away for the 4 hospitals. As can be seen, in all these graphs there is a overlap of the distributions for those new-borns that survived and those that passed away. From the analysis of these graphs, one can verify that for the indexes CRIB, SNAP, SNAP-PE and TISS, high values tend to indicate the occurrence of death. The variable WEIGHT_G has an opposite distribution, that is, an high value increases the chances of survival.









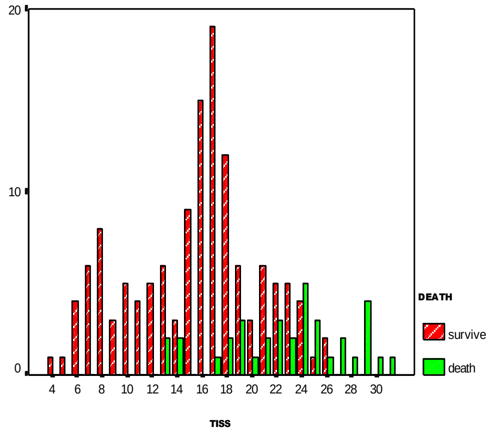


Fig. 2(a) - Frequency distributions for the new-borns that survived and for those that passed away for the 5 indexes.

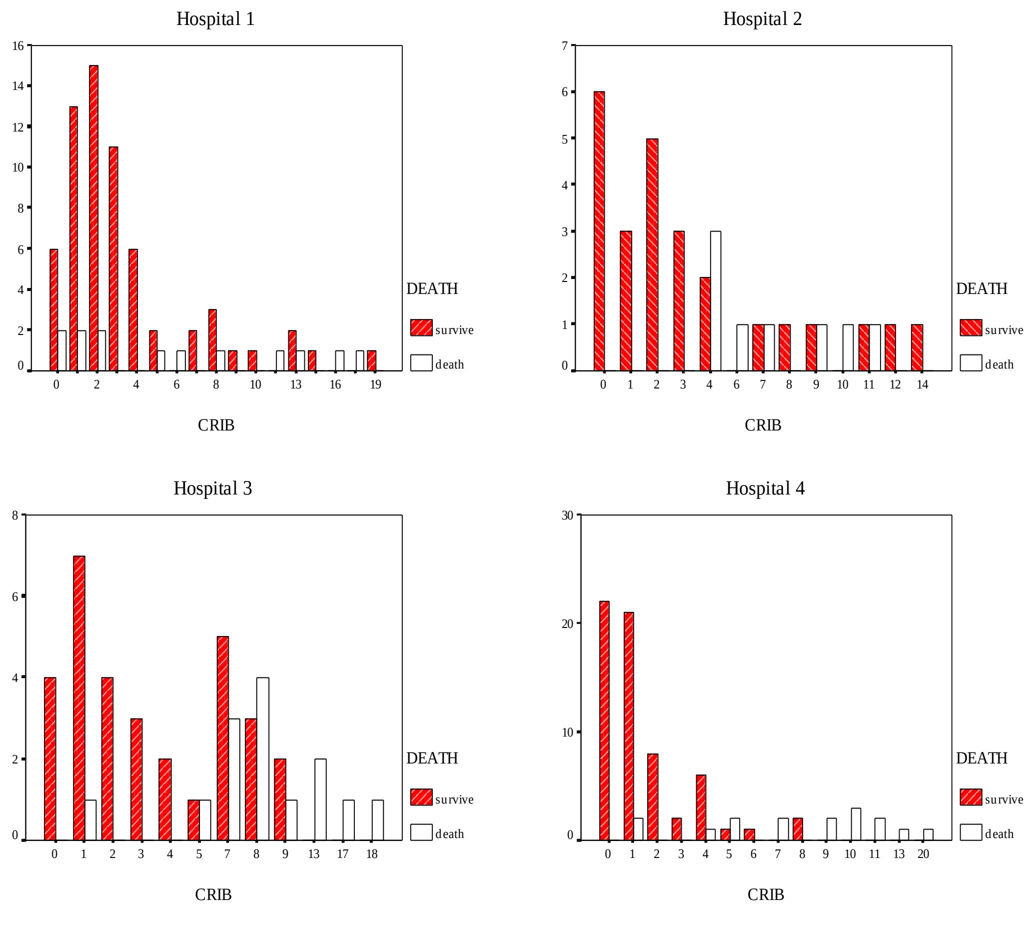


Fig. 2(b) - Frequency distributions for the new-borns that survived and for those that passed away for the 4 hospitals.

The areas under the ROC curves and standard errors

The values for the area under the ROC curve for the various indexes and hospitals were obtained from the Wilcoxon-Mann-Whitney statistics, using the SPSS package [4].The observed values and the corresponding standard errors are presented on Table 1(a) and Table 1(b), respectively.

Table 1(a) - The observed values for the area under the ROC curve for the various indexes and the corresponding standard errors

Index	**Area under ROC curve**	**Standard Error SE**
CRIB	0. 90	0.03
WEIGHT_G	0.77	0.05
SNAP	0.88	0.03
SNAP-PE	0.88	0.03
TISS	0.84	0.04

Table 1(b) - The observed values for the area under the ROC curve for the various hospitals and the corresponding standard errors

Hospital	**Area**	**Standard Error**
$\mathbf{H_1}$	0.59	0.09
$\mathbf{H_2}$	0.79	0.10
$\mathbf{H_3}$	0.84	0.07
$\mathbf{H_4}$	0.92	0.05

ROC Curves

Graphically, the ROC curve represents the probability of a true positive as a function of the probability of a false positive for a range of cut-off values. The ROC curves for the 5 indexes and for the 4 hospitals are shown in Figure 3(a) and 3 (b), respectively.

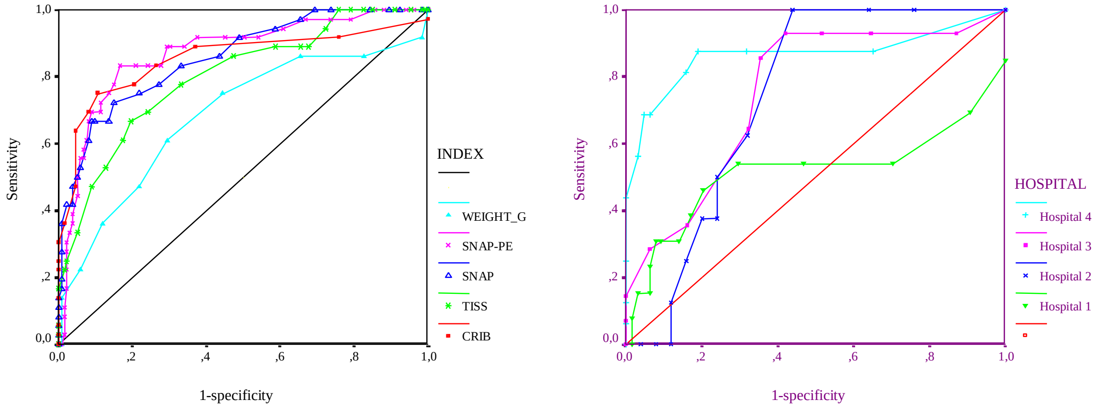


Fig. 3 (a) - The five ROC curves for the indexes; (b) –The four ROC curves for the hospitals.

From these representation, it is possible to verify that the ROC curve that is closer to the upper left corner is the one that must be preferable as predicting the risk of death in new-borns of very low weight.

Multiple comparison tests

The correlation matrixes for the deceased (r_A) and the survived (r_N) new-borns were computed using Kendal's tau from the SPSS package [4] , and are listed on Table 2.

Table 2 - Correlation matrixes for the deceased (r_A) (a), and the survived new-borns(r_N) (b).

(a) $\mathbf{r_A}$

CRIB	WEIGHT_G	SNAP	SNAPPE
- 0.1002			
0.3650	- 0.0534		
0.2632	- 0.3768	0.5846	
0.1690	0.0426	0.2504	0.1428

(b) $\mathbf{r_N}$

	CRIB	WEIGHT_G	SNAP	SNAPPE
WEIGHT_G	- 0.3679			
SNAP	0.4649	- 0.2027		
SNAPPE	0.3799	- 0.4056	0.5785	
TISS	0.4875	- 0.2581	0.4938	0.3750

The results from the multiple comparison tests for the 5 indexes and the results from the multiple comparison tests for the 4 hospitals are presented on Table 3(a) and 3(b), respectively, in terms of the p-value.

Table 3(a) - Results from the multiple comparison tests in terms of the p-value for the 5 indexes

p-value	CRIB	WEIGHT_G	SNAP	SNAPPE
WEIGHT_G	0.009			
SNAP	0.311	0.026		
SNAPPE	0.332	0.011	0.487	
TISS	0.079	0.148	0.147	0.158

Table 3(b) - Results from the multiple comparison tests in terms of the p-value for the 4 hospitals

p-value	H_1	H_2	H_3
H_2	0.071		
H_3	0.018	0.369	
H_4	0.001	0.128	0.158

7. DISCUSSION AND CONCLUSIONS

From Table 3(a) it can be seen that no significant differences were found between the diferent indexes. However, the comparison of the areas under the ROC curves points CRIB as the best index for evaluating the risk of death for new-borns with very low weight, since this presents the largest area (A=0.90) and the smallest standard error (SE(A)=0.03).

It must be pointed out that the retrospective design from the study might have affected the performance of the indexes SNAP, SNAP-PE and TISS. In routine practice, not all the tests included in these indexes are executed, and therefore, its inclusion for the study might not be justified.

Also, due to the nature of the study, and because the data for the different indexes are collected over the same set of individuals, independence cannot be guaranteed. This fact is taken into account through the correlation coefficients.

It has also been verified that the correlation between the various indexes is more significant for the survived new-borns than for the deceased ones; this might be due to the fact that the dimension of the sample of survived individuals (133) is quite large as compared to the sample of deceased new-borns (36).

From Table 3(b) it can be concluded that no significant differences were found between hospital H_1 and hospitals H_2 or H_3; however a significant difference between hospital H_1 and hospital H_4 was found.

From the comparison of the areas under the ROC curves for the hospitals, it can be said that hospital H_4 has better performance in terms of CRIB, since it does present the largest area (A=0.92) and the smallest standard error (SE(A)=0.05).

8. REFERENCES

[1] The International Neonatal Network , (1993), "*The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing perfomance of neonatal intensive care units",* The Lancet, **342**:193-98.

[2] Metz, C.E., (1978), "*Basic Principles of ROC Analysis"*,Seminars in Nuclear Medicine, Vol III, nº 4

[3] Hanley, J. A., McNeil, B., J. (1983), "*A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases",* Radiology, **148**, 839-843.

[4] SPSS for Windows, Release 7.5, SPSS Inc., 1997.